Research Article Vol. 14, No. 11 / 1 Nov 2023 / Biomedical Optics Express 5569
Biomedical Optics EXPRESS

Rapid azoospermia classification by stimulated Raman scattering and second harmonic generation microscopy

JIE HUANG,[1,2,†] XIAOBIN TANG,[3,†] ZHICONG CHEN,[4,†] XIAOMIN LI,[4] YONGQING ZHANG,[3] XIANGJIE HUANG,[2] DELONG ZHANG,[3,5] GENG AN,[4,6] AND HYEON JEONG LEE[2,*]

[1]*Zhejiang Polytechnic Institute, Polytechnic Institute, Zhejiang University, Hangzhou 310058, China*
[2]*College of Biomedical Engineering & Instrument Science; Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou 310058, China*
[3]*Interdisciplinary Centre for Quantum Information, Zhejiang Province Key Laboratory of Quantum Technology and Device, and Department of Physics, Zhejiang University; Hangzhou 310027, China*
[4]*Department of Obstetrics and Gynecology, Center for Reproductive Medicine; Guangdong Provincial Key Laboratory of Major Obstetric Diseases; Guangdong Provincial Clinical Research Center for Obstetrics and Gynecology; Guangdong-Hong Kong-Macao Greater Bay Area Higher Education Joint Laboratory of Maternal-Fetal Medicine; The Third Affiliated Hospital of Guangzhou Medical University; Guangzhou 510150, China*
[5]*dlzhang@zju.edu.cn*
[6]*angeng0505@outlook.com*
[†]These authors contributed equally
[*]*hjlee@zju.edu.cn*

Abstract: Disease diagnosis and classification pose significant challenges due to the limited capabilities of traditional methods to obtain molecular information with spatial distribution. Optical imaging techniques, utilizing (auto)fluorescence and nonlinear optical signals, introduce new dimensions for biomarkers exploration that can improve diagnosis and classification. Nevertheless, these signals often cover only a limited number of species, impeding a comprehensive assessment of the tissue microenvironment, which is crucial for effective disease diagnosis and therapy. To address this challenge, we developed a multimodal platform, termed stimulated Raman scattering and second harmonic generation microscopy (SRASH), capable of simultaneously providing both chemical bonds and structural information of tissues. Applying SRASH imaging to azoospermia patient samples, we successfully identified lipids, protein, and collagen contrasts, unveiling molecular and structural signatures for non-obstructive azoospermia. This achievement is facilitated by LiteBlendNet-Dx (LBNet-Dx), our diagnostic algorithm, which achieved an outstanding 100% sample-level accuracy in classifying azoospermia, surpassing conventional imaging modalities. As a label-free technique, SRASH imaging eliminates the requirement for sample pre-treatment, demonstrating great potential for clinical translation and enabling molecular imaging-based diagnosis and therapy.

© 2023 Optica Publishing Group under the terms of the Optica Open Access Publishing Agreement

1. Introduction

Effective disease diagnosis and classification are crucial first steps toward successful treatment, representing a significant challenge in biology and medicine. Traditional medical imaging modalities, such as magnetic resonance imaging, positron emission tomography, and ultrasound imaging, often lack the necessary combination of high spatial resolution and molecular information provided by optical imaging methods. While fluorescence microscopy has been widely used, its potential is constrained by the limited number of simultaneous targets it can detect. Additionally, specific optical modalities, for example, Raman spectroscopy, second harmonic imaging, and

#501623 https://doi.org/10.1364/BOE.501623
Journal © 2023 Received 25 Jul 2023; revised 26 Sep 2023; accepted 26 Sep 2023; published 5 Oct 2023

Biomedical Optics EXPRESS

transient absorption, provide information for lipids/proteins/nucleic acids, collagens, and pigments, respectively. However, relying solely on single-modality imaging restricts comprehensive assessment of the tissue microenvironment, which is crucial for accurate disease diagnosis and effective therapy.

This challenge is particularly evident in diagnosing azoospermia, a significant contributor to global infertility. Azoospermia, defined as the complete absence of sperm in the semen, affects millions of families worldwide and requires accurate diagnosis for proper treatment. Non-obstructive azoospermia (NOA), the most difficult to identify, represents 10%-15% of male infertility cases [1,2]. Current treatment methods such as microdissection testicular sperm extraction (micro-TESE) [3–5] only work in 50% of NOA patients due to the challenges in differentiating between normal and abnormal seminiferous tubules solely based on their color and size [6,7]. Testicular biopsy can further enhance the precision; still, the current procedure involves time-consuming hematoxylin and eosin (H&E) staining and requires highly skilled technicians and clinicians for morphology-based detection [8].

Although molecular-level information offers a more precise snapshot of cellular or tissue physiology [9], ensemble-averaged measurements from complex tissues are subject to high variability. Therefore, *in situ* molecular imaging is advantageous for clinical diagnosis and treatment monitoring. In animal models, multiphoton microscopy has been utilized to mediate sperm extraction during micro-TESE surgery by performing real-time staging of spermatogenesis based on the autofluorescence spectroscopic characteristics of sperm and Sertoli cells, and second harmonic generation (SHG) signal from collagen [10]. In addition, three-channel optical imaging, i.e., SHG, short- and long-wavelength autofluorescence, has been used to examine human testicular biopsy tissue [11]. These studies demonstrate the potential of multimodal imaging for improving surgical treatment outcomes in men with NOA.

Advancements in spectroscopy modalities hold promise for improving disease diagnosis. Raman spectroscopy, in particular, has emerged as a valuable tool for characterizing biomolecules based on their chemical bond composition and molecular structures. By utilizing the sensitivity of Raman spectroscopy to molecular bond structures, comprehensive insights into complex diseases can be obtained. Previous studies have demonstrated the potential of Raman spectroscopy in identifying both complete and incomplete spermatogenesis in seminiferous tubules from partial Sertoli-cell-only (SCO) rat models and human samples [12,13]. The abundance of spectral peaks in Raman spectroscopy contributes substantial information for disease diagnosis, highlighting its potential for effective diagnostic applications. However, the precision of single-point measurements in detecting highly heterogeneous tissues is limited, necessitating the development of an imaging technique that provides high spatial resolution with rich molecular information.

To tackle these challenges, we develop stimulated Raman scattering and second harmonic generation imaging (SRASH), which offers multidimensional molecular and structural information of tissues and a novel set of neural networks to enhance its capability for rapid disease classification (Fig. 1). The SRASH platform utilizes stimulated Raman scattering (SRS) imaging to provide fast, label-free imaging of lipids and total proteins, and SHG imaging to track collagen fibers. This method takes advantage of the intrinsic signals from intact, unstained tissue slices, eliminating the need for time-consuming sample pre-treatment before imaging. In addition, we have developed a novel algorithm, LiteBlendNet-Dx (LBNet-Dx), to extract information from the rich, high-dimensional dataset effectively. Current deep learning models, such as ResNet-50, Inception-ResNet-v2, and Swin-T [14–16], are designed for conventional RGB images, requiring large datasets to achieve accurate predictions. Here, with the new dimension of information provided by chemical imaging, LiteBlendNet leverages multi-scale feature fusion and dilated convolutions to provide an enlarged receptive field, while maintaining a lightweight structure. Furthermore, unlike previously reported margin detection methods that identify disease-related

Biomedical Optics EXPRESS

cores [17–19], NOA subtyping necessitates phenotyping of the sample, achieved through a voting algorithm (Dx) that employs weighted summation of the classification results of sample patches. Consequently, as shown in the result section, LBNet-Dx exhibits a remarkable 96.2% accuracy at the patch-level and a perfect accuracy of 100% at the sample-level in the normal and NOA diagnostic tasks. On NOA subtyping task, LBNet-Dx achieves a patch-level accuracy of 96.1% and maintains a sample-level accuracy of 100% for both NOA subtypes, showcasing unprecedented performance.

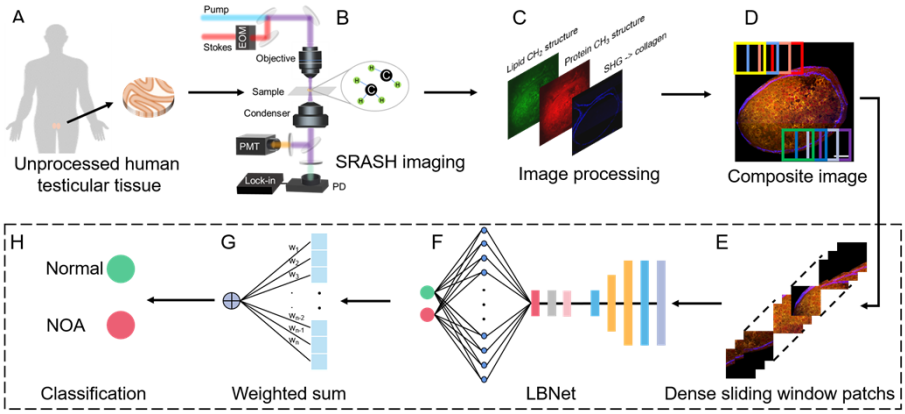

Fig. 1. Concept of SRASH. (A) The clinical unprocessed human testicular tissue samples were collected and frozen sections were made. (B) Illustration of SRASH microscope setup. Pump and Stokes refer to excitation lasers. EOM: electro-optic modulator; PMT: photo-multiplier tube; PD: photodiode. (C) The lipid (2850 cm^{-1}) channel, protein (2930 cm^{-1}) channel, and collagen (SHG) channel were colored in green, red, and blue, respectively. (D) Composite image was generated by merging three channel images. (E-H) Structure of LBNet-Dx, containing patch generation module, LBNet for patch-level classification, weighted (weight w_1 to w_n) sum module for decision, and final sample-level classification probability.

2. Methods

2.1. Patients and tissue specimens

The testicular tissues were obtained from patients who went through sperm retrieval surgery at The Third Affiliated Hospital of Guangzhou Medical University. A total of 31 patient samples were collected for this study, in which 17 patients were diagnosed with obstructive azoospermia (OA) or anejaculation who have received testicular sperm aspiration (TESA) for assisted reproductive therapy and 14 NOA patients who have received micro-TESE. All tissues were snap-frozen in liquid nitrogen immediately after the tissue retrieval for frozen sectioning. 31 tissue slides were prepared, which were verified and graded following the World Health Organization guidelines by at least two pathologists. Written informed consent was obtained from all subjects, and all experimental protocols were approved by the ethics committee at The Third Affiliated Hospital of Guangzhou Medical University.

2.2. SRASH microscopy system setup

SRASH microscope is illustrated in Fig. S1. The fundamental 1031 nm beam (~2ps) was used as the Stokes, and the wavelength tunable output (700-990 nm) was used as the pump. The pump beam was tuned to 795.9 nm and 790.9 nm for the two Raman bands at 2850 cm^{-1} (lipid) and 2930 cm^{-1} (protein). The Stokes beam was intensity modulated by an electro-optical modulator

Biomedical Optics EXPRESS

at 80 MHz and collinearly combined with the pump beam through a dichroic mirror (650 nm cutoff, Thorlabs). The combined beam was delivered to the laser scanning microscope (BX51WI, Olympus) and focused onto the samples with an objective (UPLSAPO 60XWIR, NA 1.2 water, Olympus). The pump beam was detected by a homemade photodiode, then the SRS signal was extracted by a lock-in amplifier. SHG signals from collagen fibers were collected through the same objective and detected with a photomultiplier tube. Each field of view (FOV) was imaged with a size of 1200×1200 pixel^2 (360×360 μm^2) with a lateral resolution of ~ 300 nm, and 10 μs pixel dwell time. Each FOV was averaged three times.

2.3. Quantitative and statistical analysis

The raw images were subject to background subtraction and laser power normalization. To analyze and compare lipid droplet content and size, a threshold was applied on the 2850 cm^{-1} channel to extract the lipid droplet feature for quantification. The sample sizes for quantitative and statistical analysis were normal (n = 34); NOA (n = 27); SPT (n = 14); SPG (n = 13), after excluding outlier data through interquartile range (IQR) calculation. The uncommon NOA subtypes (SCO tubule and empty seminiferous tubule) was not included in the quantitative analysis. For statistical analysis, data were first tested for normality. If the data is non-Gaussian, a nonparametric Mann–Whitney U-test was performed. Otherwise, a one-tailed student's t-test was performed. $p < 0.05$ was judged to be statistically significant.

2.4. Image preprocessing and data augmentation

All image processing steps (Fig. S2) were done using ImageJ. First, the image data set was generated by converting the raw SRASH images into three-channel RGB images, where red, green, and blue represent 2930 cm^{-1}, 2850 cm^{-1}, and SHG channels, respectively. In the normal and NOA diagnostic experiments, the image dataset was divided into a training/validation set (54 images, 26 normal and 28 NOA) and a test set (22 images, 12 normal and 10 NOA) using a 7:3 ratio, ensuring a balanced number of patches in both classes. For the diagnostic experiments involving the two subtypes of NOA (not including a small number of uncommon NOA subtypes), we selected 15 SPT and 19 SPG images from each of the 38 NOA images. Out of these, 3 SPT and 5 SPG images were used as the test set, while the remaining images were assigned to the training set. A large number of patches were obtained by the dense sliding window algorithm. The sliding window size is 300×300 pixels, and each step is shifted by 100 pixels to generate image blocks from the RGB images of 1200×1200 pixels. Through this process, 100 patches were generated per images initially. Patches with more than 90% of background pixels (tissue-free area with the pixel value set as 0) were determined to be invalid patches and be removed, resulting in 30 - 90 valid patches per images (Fig. S3). The valid patches were subject to data augmentation, including random horizontal/vertical flips, random rotation of 45 degrees on these patches, and brightness/contrast/saturation jitters with a magnitude of 0.4 using the PyTorch framework, to enhance the generalization and robustness. Finally, these patches were fed into LBNet-Dx for training after z-score normalization using means of (0.485, 0.456, 0.406) and standard deviations of (0.229, 0.224, 0.225) for the three channels.

2.5. Hyperparameter optimization by cross-validation

The hyperparameters of LBNet were optimized using a 9-fold cross-validation approach. First, the 54 images from the training/validation set were randomly divided into 9 groups (9 folds) of equal size. Then, during each iteration of the cross-validation process, one fold was held out as a validation set while the model was trained on the remaining 8 folds. This process was repeated 9 times, with a different fold used as the validation set in each iteration, resulting in a set of optimized hyperparameters. The average cross-validation accuracy achieved by the model was 96.1%, with a variance of 4e-4. Finally, the optimized network was trained on the entire training

Biomedical Optics EXPRESS

set of 54 images and evaluated on the remaining 22 images. Due to the limited sample size available for NOA subtype diagnosis, we did not employ cross-validation specifically for this issue but directly utilized the hyperparameters determined from the aforementioned process.

2.6. LiteBlendNet implementation

LiteBlendNet (Fig. S4) is a lightweight deep learning model implemented using PyTorch (v1.10) with a compact architecture, consisting of just 3 million parameters. The input image undergoes initial feature extraction through several convolutional layers and batch normalization layers. The extracted feature maps are then passed through the BlendModule, a feature fusion module that performs multi-scale feature fusion. The BlendModule comprises three branches, each consisting of a combination of standard convolutional layers and dilated convolutions, allowing for feature extraction from different receptive fields. Additionally, LiteBlendNet incorporates residual connections, enabling direct feature propagation between layers to facilitate information flow and mitigate the vanishing gradient problem. The feature maps are subsequently down-sampled using pooling layers to reduce their spatial dimensions. Following that, a sequence of convolutional layers, SiLU activation functions, and batch normalization layers further process the feature maps. Finally, a global average pooling layer is applied to obtain a fixed-length feature vector, which is then fed into a fully connected layer to produce the final classification output. The categorical cross-entropy loss function was used as the loss function in the gradient descent process. The hyperparameters of the network were determined as described in the cross-validation experiments. For the training, the AdamW optimizer with an initial learning rate of 8e-4, $\beta 1 = 0.6$, $\beta 2 = 0.999$, ε of 10-8 (numerical stability constant), and batch size = 40 was set, and the learning rate was set to decay by 20% every 10 epochs. Regarding the discrimination between normal and NOA, our final report is based on the results obtained after 80 training epochs. For distinguishing between the two subtypes of NOA, our analysis relies on the outcomes achieved after 50 epochs. The network was trained and tested on a server (Intel Xeon Gold 6248 CPU, Tesla V100 PCIe 32GB GPU).

2.7. Inference algorithm for sample-level diagnosis

The inference algorithm for sample-level diagnosis involves mapping predictions made at the patch-level to the sample-level classification in order to produce the final diagnosis result. Typically, the categories of a sample are determined by combining the classification results of all patches that belong to that sample. However, it is important to recognize that each patch may contain different amounts of useful information when fed into the LBNet. Thus, to calculate the categories of a sample, the patches must be weighted according to their effective information. To determine the weights, the number of valid pixels contained in each channel of the patch was calculated, and this information is used to perform a weighted summation of the probability distributions of the patch outputs. Then, the probability was renormalized to obtain the sample-level probability distributions. To ensure that the classification method is reliable at the sample-level, a threshold requirement was implemented. Specifically, if the probability of a class is greater than 80%, the classification result is considered valid; otherwise, it is marked as unidentifiable. This requirement ensures that the classification results have a high level of confidence (See Fig. S5 for details).

2.8. Implementation of machine learning methods

Widely used machine learning algorithms, including Support vector machines (SVM), K-Nearest Neighbor (KNN), and Decision Trees (DTs), were selected for performance comparison with LBNet-Dx. In addition, Gradient Tree Boosting (GTB) was also compared as an ensemble learning method that enhances multiple weak classifiers into one strong classifier. These algorithms are implemented by calling the Scikit-Learn library:

Biomedical Optics EXPRESS

SVM: sklearn.svm.SCV
KNN: sklearn.Ineighbors.KNeighborsClassifier
DTs: sklean.tree.DecisionTreeClassifier
GTB: sklearn.ensemble.GradientBoostingClassifier

Same input and output processing procedures as the LBNet-Dx were performed for these machine learning methods.

3. Results

3.1. SRASH imaging of testicular tissues

Multimodal SRASH images were acquired on frozen tissue slices from 31 patients (Table S1). Firstly, the normal testicular tissues from OA patients were imaged to evaluate the capability of SRASH imaging to reveal the structure of seminiferous tubules (Fig. 2(A-D)). A significant difference in the contents and distribution of lipids and proteins was observed at the subcellular level (Fig. 2(A-B)). As ubiquitously found in cells and tissue, protein component provided a clear tissue structure and morphological features (Fig. 2(B)). Specifically, less lipid found in the nucleus compared to cytoplasm provided cellular morphology information comparable to the H&E result but without labeling. Besides, the lamina propria of the seminiferous tubule was highlighted by showing only a protein signal with a strong SHG signal (Fig. 2(C)), contributing from collagen fibers. Importantly, due to the minimum sample preparation, SRASH imaging better retained the tissue intact than H&E slides where cells or biomolecules were lost in the washing and staining procedures (Fig. 2(D) and Fig. S6). As a result, several small lipid droplets (LDs) were observed (Fig. 2(E), white arrows). Overall, SRASH demonstrated label-free, molecular-selective imaging of human testicular tissues showing highly heterogenic molecular distributions within a single seminiferous tubule.

To investigate the spatial molecular signatures of abnormal seminiferous tubules, we performed SRASH imaging on the testicular tissue slice from the NOA patients. According to the maturity of seminiferous tubules, the tissues from NOA patients were divided into two common NOA subtypes (Table S1), blocked in sperm cells (SPT), blocked in spermatogonia (SPG); the remaining were categorized as uncommon NOA subtypes (SCO tubule and empty seminiferous tubule). From the SRASH images of NOA tissues, several molecular signatures were identified (Fig. 2(F-I) and Fig. S7). The change in the lamina propria was observed. In both types of NOA tissues, collagen fibers were thickened and often truncated (Fig. S7(C) and Fig. S7(F)), which is even more obvious in SPG (Fig. S7(F)). Interestingly, compared to the normal tissue (Fig. 2(A)), lower lipid signals were found in both types of NOA tissues, with much smaller LDs scattered in the seminiferous tubules (Fig. S7(A) and Fig. S7(D)). While only a few LDs were found in SPT (Fig. 2(F-G)), almost no LDs were found in SPG (Fig. 2(H-I)). In SPT seminiferous tubules, some clean protein droplets (Fig. 2(G), blue arrows) were observed in addition to some LDs (Fig. 2(G), white arrows), while only LDs can be observed in normal and SPT groups. In uncommon NOA subtypes, such as SCO tubule and empty seminiferous tubule, SRASH images also provided distinct diagnostic features (Fig. S8). In the SCO tubule (Fig. S8(A-C)), obvious vitreous degeneration occurred in the seminiferous tubules, with a few LDs. The vacuity of the tubule lumen, thickening of lamina propria, and significant hyperplasia of interstitial fibrous tissue outside the tubule could all be significantly observed (Fig. S8(D-E)), which were consistent with H&E (Fig. S8(F)).

For quantitative analysis, two indexes reflecting the amount of lipid accumulation were compared: the mean lipid intensity of LD and the area percentage of LD in seminiferous tubules. The mean intensity of LD increased significantly (Fig. 2(J)), and the area decreased significantly in the NOA group compared to the OA group (Fig. 2(K)). Such a pattern indicated that although more LDs are presented in normal tissues from OA patients, they are denser in NOA patients. No significant difference was observed between the two NOA subgroups in the mean intensity

Biomedical Optics EXPRESS

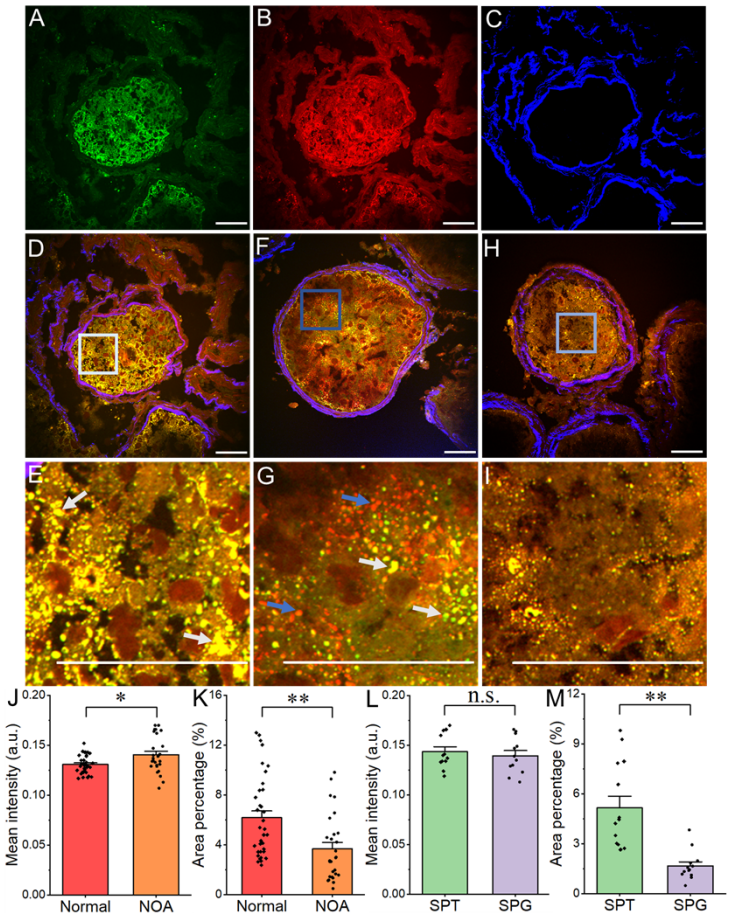

Fig. 2. SRASH images of seminiferous tubules and quantitative analysis of the lipid distribution. (A-C) Raw SRASH images of normal testicular. (A) Raw SRS image at 2850 cm^{-1} (lipid). (B) Raw SRS image at 2930 cm^{-1} (protein). (C) Raw SHG image (collagen). (D) Composite image of lipid (green), protein (red), and collagen (blue) channels. (E) Corresponding zoomed images from rectangles in (D), LDs (white arrows). (F) Three-channel composite image of SPT seminiferous tubule. (G) Corresponding zoomed images from rectangles in (F), protein droplets (blue arrows), LDs (white arrows). (H) Three-channel composite image of SPG seminiferous tubule. (I) Corresponding zoomed images from rectangles in (H). (J-M) Quantitative analysis of the lipid distribution in the seminiferous tubules. Each point represents one seminiferous tubule. Normal (n = 34); NOA (n = 27); SPT (n = 14); SPG (n = 13). Data are shown as mean ± SEM. *P < 0.05, **P < 0.01, n.s. non-significant. Scale bar: 50 μm.

Biomedical Optics EXPRESS

(Fig. 2(L)), but the area percentage of LDs in SPG decreased significantly (Fig. 2(M)). Given that these two subgroups of NOA were divided based on their spermatogenic potentials, these results suggest the correlation between LDs and spermatogenesis: fewer number of LDs, the weaker ability of spermatogenesis.

3.2. Deep learning-based inference algorithm for assisted diagnosis

To enhance the efficiency of diagnosing azoospermia, we developed a deep-learning-based inference algorithm, LBNet-Dx, and evaluated its performance on a set of 76 images. We note that the sample size is relatively small for the deep learning, which tend to have model overfitting problem. To address challenge, we designed a lightweight model and employed approaches, such as data augmentation, L1 regularization, and early stopping. Furthermore, we performed a 9-fold cross-validation [20] to demonstrate the reliability of LBNet on the small sample size (see Methods). The average accuracy achieved in the 9-fold cross-validation was 96.1%, with a variance of 4e-4, demonstrating the high precision and stability of the algorithm.

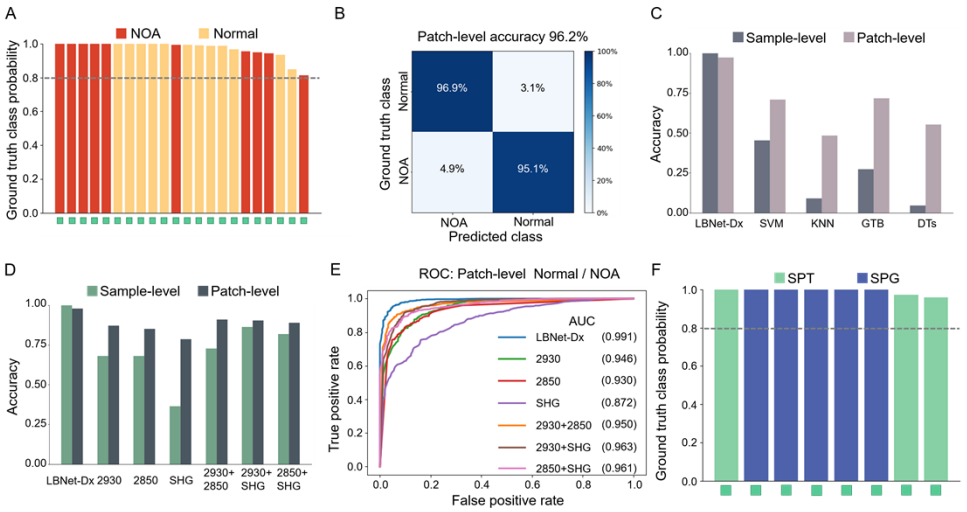

Fig. 3. Comparison of LBNet-Dx classification of SRASH imaging. (A) The classification probabilities of the test set (n = 22) on the ground truth classes are plotted in descending order, with green indicating correct classification. Dashed line indicates classification threshold. (B) Patch-level confusion matrix on the test set. (C) Accuracy comparison of LBNet-Dx and conventional machine learning methods. (D) Accuracy comparison between different channel combinations. (E) ROC curve comparison of patch level between different channel combinations. The content in brackets in the legend is the AUC value of the curve. (F) The classification probabilities of the subtypes of NOA test set (n = 8) on the ground truth classes are plotted in descending order, with green indicating correct classification. Dashed line indicates classification threshold.

In the subsequent calculation of the patch classification categories using a probability distribution, we determined the optimal threshold value of 0.393 (Fig. S9(A)) based on Youden's index from the training results. Patches with a positive probability higher than the threshold value were classified as the NOA group; otherwise, they were considered as a normal group. To determine the final category of each sample, we obtained the sample-level probability distribution by weighting the sum and renormalizing results from all patches within the sample (Fig. S10). Additionally, to ensure the reliability of the classification results, we set a probability threshold of 0.8. Any probability below this threshold was considered not reliable and marked as unidentifiable. LBNet-Dx exhibited nearly perfect performance on the test dataset. At the sample-level, the

Biomedical Optics EXPRESS

classification accuracy achieved is 100% across all 22 samples (Fig. 3(A)). The patch-level classification accuracy was 96.2%, as evident from the confusion matrix (Fig. 3(B)), and the area under curve (AUC) value on the receiver operating characteristic (ROC) curve was 99.1% (Fig. 3(E)). These results indicate that LBNet-Dx performs on par with a perfect classifier.

In comparison to several conventional machine learning models, including support vector machines (SVM), K-nearest neighbors (KNN), gradient tree boosting (GTB), and decision trees (DTs), LBNet-Dx outperformed all of them significantly at both the sample and patch levels (Fig. 3(C), Table 1). Specifically, at the patch level, LBNet-Dx achieved a recall of 96.9%, specificity of 95.1%, precision of 96.5%, and F1 value of 0.967. In contrast, traditional machine learning methods, such as KNN and DTs only reached ~50% accuracy, while SVM and GTB achieved 70% classification accuracy, which is insufficient for two-classification problems (see Supporting Material for implementation of machine learning). At the sample-level, none of the tested machine learning algorithms achieved more than 50% accuracy after the weighted summation process and threshold requirement. Importantly, we explored the relationship between patch-level and sample-level accuracies in the inference algorithm through 1-million-time simulations (Fig. S11), which revealed the critical need for extremely high patch-level accuracy to ensure reliable inference. We also compared the performance of LBNet-Dx with other reported deep learning models, and found better performance using LBNet-Dx for classification at both patch-level and sample-level (Table S2). From this perspective, LBNet-Dx performs exceptionally well and more reliably than the traditional machine learning and deep learning models we tested.

Table 1. Comparison between LBNet-Dx and conventional machine learning methods

Patch-level	LBNet-Dx	SVM	KNN	GTB	DTs
Accuracy	0.962	0.710	0.484	0.718	0.553
Recall	0.969	0.784	0.279	0.770	0.706
Specificity	0.951	0.606	0.772	0.645	0.340
Precision	0.965	0.736	0.632	0.753	0.600
F1	0.967	0.759	0.387	0.761	0.648

Theoretically, the multi-channel information in SRASH imaging is essential for accurate diagnosis using LBNet-Dx. To confirm this, we compared results using single- or two-channel images. As shown in Fig. 3(D), all three channels in SRASH imaging are required for LBNet-Dx to achieve the highest accuracy. Notably, while the SHG channel alone had the lowest prediction accuracy, its information was critical for near 100% accuracy in the three-channel group, as indicated by the significant improvement of the ROC curves compared to the two-channel group containing only SRS images at 2850 and 2930 cm^{-1} (Fig. 3(E)). We also demonstrated the performance of LBNet-Dx without 2850 or 2930 cm^{-1} channel (Fig. 3(D- E)). The results showed a noticeable decrease in accuracy at both the patch and sample level. Overall, these results highlight the essential contribution of multi-channel information obtained from SRASH to the classification performance of LBNet-Dx algorithm.

To further demonstrate the potential of LBNet-Dx in classifying NOA subtypes, we performed classification on 15 SPT and 19 SPG images (see Methods). Even with such a limited amount of data, LBNet-Dx achieved a patch-level classification accuracy of 96.1% and a sample-level classification accuracy of 100% (Fig. 3(F)). We note that small sample size introduces uncertainties in evaluating the performance of the model. The potential overfitting was mitigated by applying various regularization methods. Still, there was a noticeable discrepancy in the AUC performance between the training and test sets (Fig. S9), suggesting limitations in the generalization ability of the model. Nevertheless, the overall performance of LBNet-Dx on the NOA subtyping has demonstrated the potential of this model for performing disease subtyping.

Biomedical Optics EXPRESS

3.3. Weight localization of LBNet-Dx prediction

Unlike typical end-to-end deep learning approaches that follow a "black box" method, we demonstrated the validity of LBNet-Dx algorithm by image-wise weight localization. It is important to recognize that not all pixels have equal importance in the decision-making process of the algorithm. Therefore, identifying "hot spots" for classification can aid clinicians in accurately and efficiently identifying relevant regions in the tissue samples for staging or phenotyping. Furthermore, recognizing disease-related spatial features can provide valuable insights into underlying mechanisms. To accomplish this goal, we utilized Gradient-Weighted Class Activation Mapping (Grad-CAM) [21] to highlight the most critical regions of the images used by LBNet-Dx in its decision-making process. Grad-CAM calculates the rate of change of the final convolutional layer of the network, which captures the most complex and high-level features of the image, to the predicted class score. Subsequently, each activation map in the final convolutional layer is multiplied by the corresponding gradient value, and the resulting weighted feature maps are summed to generate a visualization showing the image regions that are most discriminatory for classification.

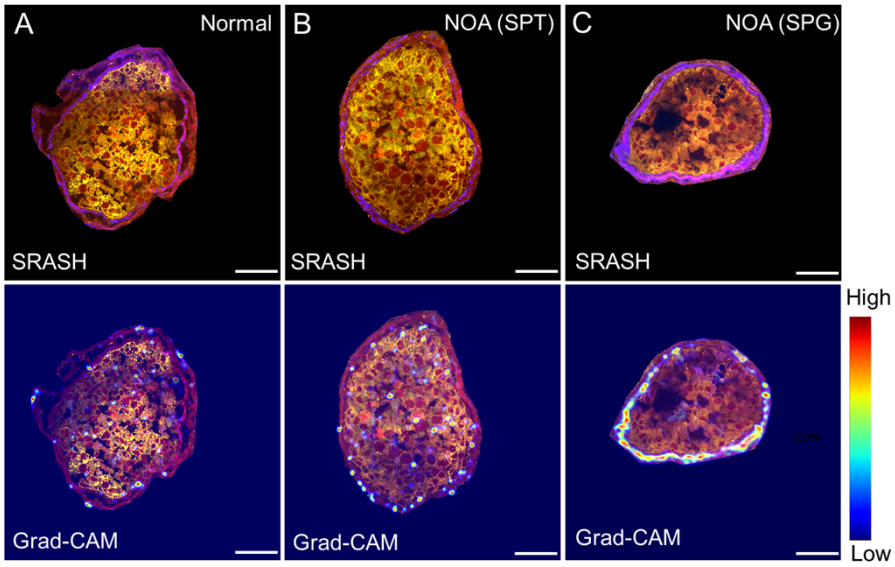

Fig. 4. Weight localization of LBNet-Dx processes. Composite Grad-CAM localization maps of Normal (A) and NOA (B-C) samples in the test set. The red and blue colors of the heatmap correspond to high and low probabilities of abnormal spermatogenesis. Scale bar: 50 µm.

Weight localization analysis revealed that LBNet-Dx accurately identified valid pixels, as evidenced by the hot spots in patches from the tissue edge (Fig. 4). Regions predicted with a higher probability of abnormal spermatogenesis are colored red in the heatmaps, whereas those with a lower probability are colored blue. Importantly, we observed a correlation between the distribution of hot spots and the SHG signal and LD aggregation. While the hot spots in both normal and NOA tissue areas were not concentrated on LD-rich areas (Fig. 4), SPG samples (Fig. 4(C)) received more attention in areas with higher SHG signals than normal (Fig. 4(A)) and SPT (Fig. 4(B)). This suggests that lipid-rich areas are less relevant to abnormal spermatogenesis, while collagen fibers play a more significant role. These results affirm that the classification made by SRASH is closely linked to biomolecule-dependent signals, highlighting the potential for identifying disease-related features for diagnosis and mechanism studies.

Biomedical Optics EXPRESS

4. Discussion

We have developed a robust approach to identify NOA without tissue staining or labeling by the SRASH platform. Compared to conventional histological evaluation, which requires extensive washing and processing, SRASH imaging requires minimal sample preparation and preserves more intact tissue for analysis. In addition, by visualizing the spatial distribution of multiple biomolecules such as lipids, proteins, and collagen fibers in testicular tissue, SRASH imaging sheds light on the underlying mechanisms of azoospermia, offering new possibilities for disease diagnosis and treatment.

It is important to note that multi-species information obtained from the SRASH imaging is essential for final classification. In particular, NOA patients exhibited increased collagen fiber deposition in the thickened lamina propria and reduced LD in the seminiferous tubules compared to normal tissue from OA patients. Grad-CAM analysis supports the importance of the lamina propria as a critical feature in distinguishing normal and NOA seminiferous tubules, suggesting that alteration in collagen fiber structure may be a signature of the disease. As previous studies have shown that the thickness of the lamina propria is negatively correlated with spermiogenesis [22,23], collagen fiber deposition may play an important role in the development of NOA. Although the collagen fiber information from the SHG signal alone was insufficient for the sample classification, this channel was required to reach highly accurate prediction. Similarly, the combination of SRS and SHG has shown great promise in imaging complicated tissue samples for studying diseases [24–26]. These results further emphasize the importance of multimodal imaging provided by the SRASH platform, and combining multiphoton fluorescence platform may obtain more dimensional information, further improve performance for spermatogenesis studies in the future, such as the treatment of azoospermia, exploring the pathogenesis of NOA.

Furthermore, SRASH imaging quantitatively measures metabolic status with subcellular resolution, providing lipid distribution in single seminiferous tubules. Reduced LD was found in AI tissues, which may indicate potential regulation of spermatogenesis by lipid metabolism. Indeed, lipid metabolism is closely related to spermatogenesis [27–29]; therefore, one possible mechanism is that reduction of lipid reservoir could lead to energy deprivation in Sertoli cells, which are essential for supporting germ cell membrane remodeling. Further mechanistic studies are expected to provide new insights into novel therapeutic strategies for azoospermia.

Multimodal imaging has emerged as a promising tool for effective disease diagnosis, offering a wealth of compositional information, molecular concentrations, and spatial distribution. Yet, extracting useful information from this high-dimensional data remains challenging. While conventional machine learning methods such as SVM, KNN, and DTs have been used for medical image-based disease diagnosis [30–32], recent advancements in deep learning have significantly improved medical image processing. Deep learning algorithms can automatically extract features from medical images, facilitating early disease prediction [33,34], clinical diagnosis [17,18,30,35–39], and tissue margin detection with high accuracy [19,40–42]. However, a significant challenge for machine learning is the scarcity of standardized datasets for training, particularly when advanced imaging techniques are employed. Therefore, an approach that can yield robust and accurate predictions while using a small dataset for training would be invaluable.

High-dimension data generated by SRASH imaging is utilized efficiently for clinical diagnosis with the help of the deep learning algorithm LBNet-Dx developed in this study. LBNet is a specifically designed lightweight network that mitigates the reliance on large-scale datasets, making it suitable for working with small datasets, which is crucial in the context of limited availability of samples in biomedical imaging. LBNet incorporates multi-scale feature fusion and dilated convolutions to enhance spatial information integration, resulting in superior classification performance compared to traditional classification algorithms. These characteristics contribute to its significance in the field of biomedical imaging analysis. The weighted voting scheme in the inference algorithm enhances robustness and reliability, while the threshold mechanism

Biomedical Optics EXPRESS

eliminates the risk of misdiagnosis and underdiagnosis. SRASH performs better classification accuracy and diagnosis speed than previous work using Google Cloud AutoML Vision to classify H&E images [43]. In combination with Grad-CAM analysis, SRASH could guide rapid and efficient diagnosis. In our study, we utilized LBNet-Dx for the diagnosis of two subtypes of NOA. However, due to the limited size of the dataset, the reliability of the obtained results is somewhat compromised. In future research, expanding the dataset's size will further bolster the model's performance and reliability, enabling a deeper understanding of NOA subtypes and providing more robust support for their clinical applications. With these capabilities, SRASH can be applied to diagnosing azoospermia and other diseases where subtyping or staging is needed, such as gout, mammary gland calcification, and skin diseases.

5. Conclusion

In summary, SRASH imaging combined with LBNet-Dx has the potential to serve as an efficient diagnostic tool for the accurate and rapid classification of diseases, as demonstrated in the disease classification of azoospermia and its subtypes. By enabling label-free imaging of frozen tissue slices and automatic voting, this approach can greatly improve clinical outcomes for male infertility patients, surpassing traditional azoospermia diagnostic procedures in speed. Additionally, this could facilitate the guidance of surgeons performing micro-TESE when an *in vivo* imaging scheme is adopted. Given the small sample sizes for training the algorithm, this approach is highly suitable for clinical implementation. Furthermore, it can be extended to other tissue imaging and disease diagnosis applications where multimodal imaging is employed. This technique would allow more informed decisions about the best treatment options for male infertility and potentially other diseases through precise tissue imaging and analysis.

Funding. Guangdong Science and Technology Department (202102010075); Natural Science Foundation of Guangdong Province (2019A1515011439); Fundamental Research Funds for the Central Universities (2020QNA5027); National Natural Science Foundation of China (12074339, 32050410293, 82171589, 82203709).

Disclosures. The authors declare no conflicts of interest.

Data Availability. All data generated or analyzed during this study are available upon request. The code is included in the Supplementary information file.

Supplemental document. See Supplement 1 for supporting content.

References

1. J. P. Jarow, M. A. Espeland, and L. I. Lipshultz, "Evaluation of the azoospermic patient," The Journal of urology **142**(1), 62–65 (1989).
2. P. Thonneau, S. Marchand, A. Tallec, M.-L. Ferial, B. Ducot, J. Lansac, P. Lopes, J.-M. Tabaste, and A. Spira, "Incidence and main causes of infertility in a resident population (1 850 000) of three French regions (1988–1989)," Hum. Reprod. **6**(6), 811–816 (1991).
3. P. N. Schlegel, "Testicular sperm extraction: microdissection improves sperm yield with minimal tissue excision," Hum. Reprod. **14**(1), 131–135 (1999).
4. P. Devroey, J. Liu, Z. Nagy, H. Tournaye, S. J. Silber, and A. C. Van Steirteghem, "Normal fertilization of human oocytes after testicular sperm extraction and intracytoplasmic sperm injection," Fertil. Steril. **62**(3), 639–641 (1994).
5. N. Punjani, C. Kang, R. K. Lee, M. Goldstein, and P. S. Li, "Technological advancements in male infertility microsurgery," J. Clin. Med. **10**(18), 4259 (2021).
6. R. Ramasamy, N. Yagan, and P. N. Schlegel, "Structural and functional changes to the testis after conventional versus microdissection testicular sperm extraction," Urology **65**(6), 1190–1194 (2005).
7. H. Okada, M. Dobashi, T. Yamazaki, I. Hara, M. Fujisawa, S. Arakawa, and S. Kamidono, "Conventional versus microdissection testicular sperm extraction for nonobstructive azoospermia," J Urol **168**(3), 1063–1067 (2002).
8. R. I. McLachlan, E. Rajpert-De Meyts, C. E. Hoei-Hansen, D. M. de Kretser, and N. E. Skakkebaek, "Histological evaluation of the human testis–approaches to optimizing the clinical value of the assessment: mini review," Hum Reprod **22**(1), 2–16 (2007).
9. A. Amaral, J. Castillo, J. Ramalho-Santos, and R. Oliva, "The combined human sperm proteome: cellular pathways and implications for basic and clinical science," Hum Reprod Update **20**(1), 40–62 (2014).
10. R. Ramasamy, J. Sterling, E. S. Fisher, P. S. Li, M. Jain, B. D. Robinson, M. Shevchuck, D. Huland, C. Xu, S. Mukherjee, and P. N. Schlegel, "Identification of spermatogenesis with multiphoton microscopy: an evaluation in a rodent model," J Urol **186**(6), 2487–2492 (2011).

Biomedical Optics EXPRESS

11. B. B. Najari, R. Ramasamy, J. Sterling, A. Aggarwal, S. Sheth, P. S. Li, J. M. Dubin, S. Goldenberg, M. Jain, B. D. Robinson, M. Shevchuk, D. S. Scherr, M. Goldstein, S. Mukherjee, and P. N. Schlegel, "Pilot study of the correlation of multiphoton tomography of ex vivo human testis with histology," J Urol **188**(2), 538–543 (2012).
12. E. C. Osterberg, M. A. Laudano, R. Ramasamy, J. Sterling, B. D. Robinson, M. Goldstein, P. S. Li, A. S. Haka, and P. N. Schlegel, "Identification of spermatogenesis in a rat sertoli-cell only model using Raman spectroscopy: a feasibility study," J Urol **192**(2), 607–612 (2014).
13. Y. Liu, Y. Zhu, L. Di, E. C. Osterberg, F. Liu, L. He, H. Hu, Y. Huang, P. S. Li, and Z. Li, "Raman spectroscopy as an ex vivo noninvasive approach to distinguish complete and incomplete spermatogenesis within human seminiferous tubules," Fertil. Steril. **102**(1), 54–60.e2 (2014).
14. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition* (2016), pp. 770–778.
15. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*(2017).
16. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*(2021), pp. 10012–10022.
17. T. C. Hollon, B. Pandian, and A. R. Adapa, *et al.*, "Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks," Nat. Med. (N. Y., NY, U. S.) **26**(1), 52–58 (2020).
18. L. Zhang, Y. Wu, B. Zheng, L. Su, Y. Chen, S. Ma, Q. Hu, X. Zou, L. Yao, Y. Yang, L. Chen, Y. Mao, Y. Chen, and M. Ji, "Rapid histology of laryngeal squamous cell carcinoma with deep-learning based stimulated Raman scattering microscopy," Theranostics **9**(9), 2541–2554 (2019).
19. J. Unger, C. Hebisch, J. E. Phipps, J. L. Lagarto, H. Kim, M. A. Darrow, R. J. Bold, and L. Marcu, "Real-time diagnosis and visualization of tumor margins in excised breast specimens using fluorescence lifetime imaging and machine learning," Biomed. Opt. Express **11**(3), 1216–1230 (2020).
20. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning* (Springer, 2013).
21. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*(2017), pp. 618–626.
22. Y. Sato, S. Nozawa, and T. Iwamoto, "Study of spermatogenesis and thickening of lamina propria in the human seminiferous tubules," Fertil. Steril. **90**(4), 1310–1312 (2008).
23. J. Volkmann, D. Muller, C. Feuerstacke, S. Kliesch, M. Bergmann, C. Muhlfeld, and R. Middendorff, "Disturbed spermatogenesis associated with thickened lamina propria of seminiferous tubules is not caused by dedifferentiation of myofibroblasts," Hum Reprod **26**(6), 1450–1461 (2011).
24. J. Ao, X. Shao, Z. Liu, Q. Liu, J. Xia, Y. Shi, L. Qi, J. Pan, and M. Ji, "Stimulated Raman scattering microscopy enables gleason scoring of prostate core needle biopsy by a convolutional neural network," Cancer Res. **83**(4), 641–651 (2023).
25. K. S. Shin, S. Men, A. Wong, C. Cobb-Bruno, E. Y. Chen, and D. Fu, "Quantitative chemical imaging of bone tissue for intraoperative and diagnostic applications," Anal. Chem. **94**(9), 3791–3799 (2022).
26. H. Jang, Z. Li, Y. Li, P. Bagheri, E. Ackerstaff, J. Koutcher, and L. Shi, "Ultrafast nonlinear multimodal metabolic imaging platform for studying aging and diseases," in *Ultrafast Nonlinear Imaging and Spectroscopy X* (SPIE 2022), pp. 27–31.
27. G. M. Oresti, J. Garcia-Lopez, M. I. Aveldano, and J. Del Mazo, "Cell-type-specific regulation of genes involved in testicular lipid metabolism: fatty acid-binding proteins, diacylglycerol acyltransferases, and perilipin 2," Reproduction **146**(5), 471–480 (2013).
28. R. Keber, D. Rozman, and S. Horvat, "Sterols in spermatogenesis and sperm maturation," J Lipid Res **54**(1), 20–33 (2013).
29. A. Lenzi, M. Picardo, L. Gandini, and F. Dondero, "Lipids of the sperm plasma membrane: from polyunsaturated fatty acids considered as markers of sperm function to possible scavenger therapy," Hum. Reprod. Update **2**(3), 246–256 (1996).
30. E. Duran-Sierra, S. Cheng, R. Cuenca, B. Ahmed, J. Ji, V. V. Yakovlev, M. Martinez, M. Al-Khalil, H. Al-Enazi, Y. L. Cheng, J. Wright, C. Busso, and J. A. Jo, "Machine-learning assisted discrimination of precancerous and cancerous from healthy oral tissue based on multispectral autofluorescence lifetime imaging endoscopy," Cancers **13**(19), 4751 (2021).
31. M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: a comprehensive review," Healthcare **10**(3), 541 (2022).
32. N. Kumar, N. Narayan Das, D. Gupta, K. Gupta, and J. Bindra, "Efficient automated disease diagnosis using machine learning models," J Healthc Eng **2021**, 1–13 (2021).
33. H. Hu, L. Gong, D. Dong, L. Zhu, M. Wang, J. He, L. Shu, Y. Cai, S. Cai, W. Su, Y. Zhong, C. Li, Y. Zhu, M. Fang, L. Zhong, X. Yang, P. Zhou, and J. Tian, "Identifying early gastric cancer under magnifying narrow-band images with deep learning: a multicenter study," Gastrointest Endosc **93**(6), 1333–1341.e3 (2021).
34. M. Y. Lu, T. Y. Chen, D. F. K. Williamson, M. Zhao, M. Shady, J. Lipkova, and F. Mahmood, "AI-based pathology predicts origins for cancers of unknown primary," Nature **594**(7861), 106–110 (2021).

Biomedical Optics EXPRESS

35. K. Aljakouch, Z. Hilal, I. Daho, M. Schuler, S. D. Krauss, H. K. Yosef, J. Dierks, A. Mosig, K. Gerwert, and S. F. El-Mashtoly, "Fast and Noninvasive Diagnosis of Cervical Cancer by Coherent Anti-Stokes Raman Scattering," Anal. Chem. **91**(21), 13900–13906 (2019).
36. Z. Liu, W. Su, J. Ao, M. Wang, Q. Jiang, J. He, H. Gao, S. Lei, J. Nie, X. Yan, X. Guo, P. Zhou, H. Hu, and M. Ji, "Instant diagnosis of gastroscopic biopsy via deep-learned single-shot femtosecond stimulated Raman histology," Nat. Commun. **13**(1), 4050 (2022).
37. X. Mei, H. C. Lee, and K. Y. Diao, *et al.*, "Artificial intelligence-enabled rapid diagnosis of patients with COVID-19," Nat. Med. **26**(8), 1224–1228 (2020).
38. Z. Song, S. Zou, and W. Zhou, *et al.*, "Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning," Nat. Commun. **11**(1), 4294 (2020).
39. L. Zhang, X. Zou, J. Huang, J. Fan, X. Sun, B. Zhang, B. Zheng, C. Guo, D. Fu, L. Yao, and M. Ji, "Label-free histology and evaluation of human pancreatic cancer with coherent nonlinear optical microscopy," Anal. Chem. **93**(46), 15550–15558 (2021).
40. R. Cao, S. D. Nelson, S. Davis, Y. Liang, Y. Luo, Y. Zhang, B. Crawford, and L. V. Wang, "Label-free intraoperative histology of bone tissue via deep-learning-assisted ultraviolet photoacoustic microscopy," Nat. Biomed. Eng **7**(2), 124–134 (2022).
41. T. Qaiser, Y. W. Tsang, D. Taniyama, N. Sakamoto, K. Nakane, D. Epstein, and N. Rajpoot, "Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features," Med. Image Anal. **55**, 1–14 (2019).
42. N. Yamato, M. Matsuya, H. Niioka, J. Miyake, and M. Hashimoto, "Nerve segmentation with deep learning from label-free endoscopic images obtained using coherent anti-Stokes Raman scattering," Biomolecules **10**(7), 1012 (2020).
43. Y. Ito, M. Unagami, F. Yamabe, Y. Mitsui, K. Nakajima, K. Nagao, and H. Kobayashi, "A method for utilizing automated machine learning for histopathological classification of testis based on Johnsen scores," Sci. Rep. **11**(1), 9962 (2021).